CHAPTER 2: TEST DEVELOPMENT, ADMINISTRATION, SCORING, AND REPORTING

Introduction

A major concern raised in our first evaluation report was whether it was feasible to develop a high quality exam within the time constraints specified in the legislation. In this chapter, we describe our review of the quality of the test forms that were administered in 2001 and also document our review of administration, scoring, and reporting procedures used for that administration.

The quality of the two test forms used in the 2001 administration is a direct result of the procedures used in developing and reviewing test questions and in selecting questions for inclusion in the first operational exam forms. We describe our review of these procedures and also discuss statistical indicators of the quality of the test questions based on data from field tests of these questions and from the operational administrations.

Once the first forms of the exam were developed, they had to be administered. For a time, it appeared that the 2001 administration would be a practice test for the students, also providing schools an opportunity to try out procedures for administering and scoring the tests. Administration of the CAHSEE created significant logistical issues for many schools. These logistical issues could, in turn, affect the quality of the examination. We provide a description of our observation of how the test was administered and some suggestions for making this process run more smoothly in the future.

A third set of issues potentially affecting test quality concerned the processing, scoring, and scaling of the tests. Issues included the care with which answer sheets were checked at the test sites and upon receipt at the scanning site, the accuracy and/or consistency of the hand scoring of the essay responses for the ELA test, and how the total scores were placed on the score scale. In May, there was the additional problem of achieving near equivalency of reported scores to those from the March administration, even though the May exam used a large number of different test questions.

The final quality issue discussed in this chapter is the reporting of the test results, both for individual students and for aggregations by school, district, county, and the state as a whole. The failure of SB 84 significantly affected the reporting of results. Initially, reports were designed for a practice test where results from each test question could be released, but passing standards would not be set so students would not be told whether they had passed or failed each test. On very short notice, the score reports had to be redesigned to include passing information. In addition, some questions had to be held secure for use in equating alternate forms, so information at the test question level was considerably more limited.

Quality of the Test Questions

The CAHSEE mathematics (math) examination consists of 80 multiple-choice questions. The English-language arts (ELA) exam consists of 58 multiple-choice reading questions, 24 multiple-choice writing questions, and 2 essay questions used to assess writing skills. Each

test question was designed to assess mastery of a specific content standard recommended by the HSEE Standards Panel and adopted by the SBE for coverage by the exam.

Professional and legal standards (e.g., those set by the American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999) require that tests, particularly those used in making important or high stakes decisions about people, be both valid and reliable. In this context, the CAHSEE is valid if it assesses the targeted content as completely as possible and does not require knowledge or skills beyond those specified in the content standards for the exam. A test is said to be reliable if it gives accurate or consistent estimates of the trait(s) being measured. One test of reliability would be, for example, if students took two (parallel) forms of the exam and achieved similar scores on both. A test cannot be valid if it gives inconsistent results, indicating it is not reliable in providing an accurate measure of the intended content. On the other hand, a test could be quite reliable, but still be invalid if it measured the wrong content. In evaluating tests, validity is the primary concern, followed by reliability as the issue next in importance.

Another key issue in professional and legal standards is fairness. Here fairness is primarily a question of whether the exam measures the targeted content in the same way for all groups of students. Note that groups may differ in mastery of the target content; in such cases, a fair test will neither overstate nor understate the extent of such differences. A test or an individual test question is "unfair" if it requires knowledge or skills beyond the targeted content that are differentially available or familiar to some groups of examinees compared with others. Test questions that are not fair by this definition are almost always also not valid because of the requirement of extraneous skills. Thus, validity as the primary concern is once again demonstrated.

The test development contractor performed a number of steps to assess all potential CAHSEE test questions for validity, reliability, and fairness. We describe these steps briefly here along with our own efforts to assess the validity, reliability, and fairness of the two forms of the exam used in the 2001 administrations.

Content, Editorial, and Sensitivity Reviews of the CAHSEE Test Questions

Each question developed or identified for use in the CAHSEE was subjected to extensive review before being tried out in a field test. Specific reviews included:

1. Editorial and content review by experienced editors on the AIR staff.

2. Content review by panels of teachers and educators familiar with the content standards.

3. Content and sensitivity reviews by subcommittees of the HSEE Standards Panel that had initially identified the targeted content standards.

4. Sensitivity review by expert panels including representation of key demographic groups.

5. Final review by CDE and SBE staff and by Board members themselves.

At each review, test questions could be flagged for revision or eliminated altogether from further consideration. We were able to observe some reviews performed by the HSEE Standards Panel and outside educators, and found them to be conducted very thoroughly. For the most part, relatively few problems were identified, suggesting that initial development and internal review processes were effective.

During the first year of our evaluation, we assembled panels of California educators and conducted an independent review of a sample of test questions. The primary question asked of each panelist was whether each test question was a fair and effective measure of mastery of the targeted content standard. Detailed results from that review were described in our Year 1 Report (Wise et al, 2000a). The general conclusion was that relatively few issues were identified and that the questions were generally of good quality. While we reviewed only a sample of CAHSEE test questions, the results suggested that the process used by CDE and the test developers to review all of the test questions was effective. This conclusion was further reinforced by the results of statistical analyses of the test questions described below.

As noted, all of the questions included in the 2001 administrations were developed by AIR and subjected to one of two tryouts or field tests. A new test contractor, the Educational Testing Service (ETS) was selected for development and administration of the CAHSEE beginning with the 2002 administration. As part of our independent evaluation, we plan to conduct a second independent review of test question quality in Spring 2002 as a check on any revisions to the development and review processes.

Statistical Analyses of the Test Questions

Test questions that had been developed or adapted during the first half of 2000 were included in the Spring 2000 Field Test. AIR reported results from that field test in August of 2000 (American Institutes for Research, 2000).

We reported our own analyses of the Spring 2000 Field Test in our June 30 and August 25 reports of that year (Wise et al., 2000a; Wise et al., 2000b). Included in those reports was an examination of the difficulty of each question (defined in terms of percent of students answering correctly). We flagged questions if they appeared to be inappropriately difficult or easy relative to other questions measuring the same standard. We also looked at whether performance on each question was consistent with performance on all of the other questions in the test (item-total correlation). This provided an indication of whether the question was effective in differentiating between high and low levels of mastery of the targeted standards. For the multiple-choice questions, we also looked at whether any of the incorrect options were selected by a significant number of high performing students as an indicator that the question might be incorrectly keyed or have multiple correct answers. For the essay questions, we examined the consistency with which independent readers scored them. We also examined a common indicator of "differential item functioning" to identify any items that were disproportionately difficult for various groups of students.

The results of the Spring 2000 Field Test indicated that a very high proportion of the questions had acceptable statistical properties and could be used in operational CAHSEE

examination forms. Nonetheless, additional test questions were needed to cover particular content standards and to support the assembly of multiple test forms.

Additional test questions were developed by AIR and included in a second field test conducted in Fall 2000. Results of that field test were reported by AIR. We have not yet had an opportunity to review AIR's documentation of the second field test, but we reported our own analyses of results from this field test in our Year 2 Report (Wise et al., 2001a). Again, relatively few questions were flagged in the review of statistical properties. More than 84% of the ELA questions and 72% of the math questions had no statistical flags at all.

Pages 14–24 of our Year 2 report (Wise et al., 2001a) show the number of test questions developed per content standard and the average percentage of students who answered these questions correctly. After reviewing the data in these tables, we concluded that there were a sufficient number of test questions to assemble at least two distinct exam forms that each covered the content standards as specified in the test plan recommended by the HSEE Standards Panel and approved by the SBE.

Our analysis of the difficulty of questions for different content standards indicated that questions assessing many of the algebra standards were disproportionately difficult (Wise et al., 2000b). Based on this finding, the CDE recommended and the SBE subsequently approved reduced coverage of algebra for the Class of 2004, while indicating an intention to increase coverage at a later time.

In comparing results from the two field tests, one interesting finding emerged that bears reporting here. In order to be able to compare statistical results from the Spring and Fall 2000 field tests, AIR included a common set of 20 multiple-choice questions in each of the four different ELA forms used in the fall field test and another common set of 20 math items in each of the four different math forms. Each of these common questions had been included in the spring field test, making it possible to compare the relative performance of students in the spring field test who were tested toward the end of 10th grade with the performance of students in the fall field test who were tested at the beginning of the 10th grade.

Table 2.1 shows the average percent of correct responses to the 20 linking items for the students in the fall field test and for students in the spring field test. For ELA, the students at the beginning of 10th grade in the fall field test actually did slightly better than the students from the spring field test who were at the end of the 10th grade. This might reflect a difference between the Class of 2002 who participated in the spring field test and the Class of 2003 included in the fall field test. The Class of 2003 may have benefited from additional instruction since the adoption of the California Content Standards.

TABLE 2.1 Comparison of Spring and Fall Performance on Linking Items

	ELA	Mathematics
Number of Linking Items	20	20
Passing Rates in Fall 2000 Field Test	Percent Correct	Percent Correct
Fall 2000 Avg. (beginning of 10th Grade)	62.8	53.0
Spring 2000 Avg. (end of 10th Grade)	61.7	57.5
Difference	-1.1	+4.5

The data in Table 2.1 show the opposite finding for mathematics. The sample of students at the beginning of 10th grade had lower rates of correct responses than the sample of students at the end of 10th grade by about 4.5 percentage points. The implication is that 10th grade course work improves student performance on the mathematics test, suggesting that many 9th graders may not yet be ready to take this exam.

We also conducted statistical analyses of student responses to the questions in the March and May 2001 operational test forms. Analyses of operational test results closely paralleled our analyses of the field test data. We examined the difficulty of each question, item-total correlations, incorrect option selection for the multiple-choice questions, consistency across scorers for the essay questions, and indicators of differential item functioning (DIF) for various examinee groups. Given the much larger sample size in the operational administration relative to the field test, we were able to examine differential functioning with much greater precision for a larger number of groups. In particular, while the number of African American students in the field test was too small to detect differential function with much precision, in the operational test data we were able to examine possible DIF for this group with much greater precision.

Preliminary results of our analyses of responses to the operational questions were reported in our Year 2 report (Wise et al., 2001). Subsequent analyses were entirely consistent with the conclusions stated in that report. A few questions were flagged for further review based on analyses of responses to the operational forms. Some simply turned out to be difficult questions and others included incorrect options that were attractive to students with partial knowledge. In no case was there any suggestion of problems that might warrant excluding the question from operational scoring.

Administering CAHSEE

The plan for administration of a practice test in Spring 2001 would also have allowed an opportunity for a dry run of test administration procedures. As described below, the joint demands of fairness and test security placed a number of difficult constraints on the administration of the CAHSEE. These constraints impacted schools and districts differently depending on the number of students tested, how student time is normally scheduled, the availability of testing space, and other factors. In this section, we describe our observations of the Spring 2001 administration and offer some suggestions for consideration in future administrations of the CAHSEE.

Sources of Information

HumRRO collected information on administration of CAHSEE from three sources:

1. Observing three schools as they administered CAHSEE

2. Monitoring training workshops for school and district personnel responsible for test coordination before the March administration and a focus group of district test coordinators after the March administration

3. Surveying a modest sample of school test coordinators

Characteristics of the test sessions observed are shown in Table 2.2. The HumRRO observer watched students take the test—attending to the pace of progress, test security, and level of distraction—and interviewed the test coordinators. While the schools varied in the ways they administered the CAHSEE, school staffs were well-prepared and provided good testing conditions. The most striking overall feature was how seriously students took the test.

TABLE 2.2 Characteristics of Schools Observed

School	Subject	School Type	Approximate Number Tested	Environment	Accommodations
A	ELA (March)	Urban	850	Classrooms	None
B	Math (March)	Rural	275	Auditorium	None
C	ELA (May)	Suburban	575	Classrooms	Special Education (Separation)

Our Spring 2001 survey of teachers and principals in the longitudinal sample of high schools included a brief survey of site coordinators. The site-coordinator survey (see Appendix C) asked for feedback on guidance received, students tested, the general approach to administering the test, and changes planned for future administrations of CAHSEE. Coordinators for 42 schools returned the survey. About half of the respondents had the title of test coordinator and another third were assistant principals.

CDE conducted a focus group with about 40 district testing coordinators between the March and May test dates to collect feedback on test logistics. The coordinators rotated through four stations to discuss issues with administering CAHSEE: (a) testing manuals, workshops, and staff development; (b) logistics, scheduling, and security; (c) test administration support; and (d) accommodation and regulations. The discussion of results from all three sources is organized by those topics.

Observations on Test Administration

Testing Manuals, Workshops, and Staff Development

The test developer and its subcontractor for processing and reporting (NCS Pearson) conducted five workshops with district and school test coordinators (HumRRO observed one of the workshops). The workshops focused on the importance of CAHSEE and the necessity for coordinators to get immersed quickly and take seriously procedures for the administration of the tests. Topics included session length, test security, and score reports. Speakers walked coordinators through the demanding requirements for receiving materials, preparing answer documents, and returning materials.

About 60% of the surveyed coordinators had read at least one of the coordinator manuals, but only half reported reading Directions for Administration. Most thought that the information in the manuals was clear, but several suggested changes, including: (a) Combine the coordinator manuals to eliminate overlap, (b) reduce restrictions on distribution of Directions for Administration, and (c) clarify the instructions for filling out the answer documents.

Feedback on workshops was also obtained via the survey delivered to the sample of high schools. About 25% of the school site coordinators in the survey had attended one of the

workshops. Although they generally felt frustrated by the uncertainties of whether the test was practice only or would count in fulfilling the new graduation requirement, the only negative comment about the content of the workshop was that not enough of it was about logistics, especially what to do with students who were not being tested.

While coordinators who attended the focus group also thought that the Directions for Administration were confusing, especially regarding the completion of background information in cases where the school had taken advantage of the precode option, they were positive about the workshops. They said that the workshops should be conducted earlier, at more sites, and with fewer people per session. One response to a question about plans for the next administration was, "Going to the conference was extremely helpful. Other site coordinators from my district did not go and they were confused. I recommended to them that they go to the meeting next time!"

CDE supported staff development through presenter workshops and teacher guides. Comments from the focus group about those efforts were strongly positive, especially for the option to access information via the Internet.

Logistics, Scheduling, and Security

Workshop participants provided feedback on issues including extended test-taking time, breaks, the length of the ELA test, and options for students not taking the test. Further consideration of these issues would be helpful.

The main logistics problem in the observed schools was balancing the option of extended time for students who needed it with test security and test conditions. Observers noted that School A did not provide extended time but had very good test security. At the end of both sessions, proctors alerted students that time was almost up and they should finish the test; they did not mention that additional time was available. Everyone took a break between the two main sections of the test. Because this school allotted more than 2 hours for each session, all students appeared to finish by the scheduled time, but some students in each session clearly rushed to complete their essays.

School B provided extended time and preserved testing conditions but did so at the cost of test security. This school tested students in an auditorium with lapboards and allowed about 3 hours for testing. (Because the school did not precode answer documents, completion of the background section took 30 minutes.) Students ignored the section breaks, moving directly to Section 2 as soon as they completed Section 1. After an hour, all students took a 13-minute break regardless of their progress on the test. After students finished Section 2, they left the auditorium. This approach traded security (students had a chance to get information on past or upcoming items during the break) for improved test conditions (by minimizing disruptions for more deliberate students). About 5% of the students had not finished by the time lunch started. They were released for lunch and told to report to a classroom to complete the test. Although this model was not typical of the schools in the survey, it was not unique: Two other schools disregarded the sections (and another plans to do so next time); five allowed students to finish the first section after the break; and six had students finish the exam after lunch.

School C tested students in classrooms but had not given proctors guidance on extended time because feedback from schools that had tested in March indicated that time was adequate. As a result proctors gave a variety of options to students who needed more time. In some classes, such students were sent to the library. In another class, students were told they could work through the break but no longer. Some students who needed time for Section 2 continued through lunch and received compensatory time for lunch. A survey respondent wrote: "When students need more time, it is a logistical nightmare."

A consistent comment from all sources was that the ELA test was too long. For example, a district coordinator commented that "kids max at 2 ½ hr," and a proctor at an observed school said, "These kids are fried." Approximately 5% of the students reported that they did not have enough time and about 9% did not attempt the final question, which was an essay. (Student response seems to contradict coordinators.) Note that plans for the 2002 administration now call for administering the ELA test over two separate days. This should ease the test length problem, but may increase security issues and also create logistical problems due to student absences on the second day.

The length of the mathematics test was not cited as a problem. Approximately 2% of the students reported lack of time as a problem and only about 1% of the students failed to answer the last question on the test. Nevertheless, district coordinators cautioned that the apparently comfortable time requirements might have been because many students who lacked algebra skills did not do those calculations.

Schools also were concerned about what to do with other students during testing. School A held a school-wide writing activity, which freed up classrooms and teachers, and gave flexibility for the lunch schedule, but also resulted in significant absenteeism. Two other schools had special school-wide activities. Focus-group coordinators reported that other schools scheduled field trips and minimum days. Most of the surveyed schools followed the regular class schedule for other students; about 25% conducted regular classes with a revised schedule. Only seven schools reported lower attendance than normal by other grades.

Focus-group discussions after 2001 testing indicate that providing meaningful instruction for classes with a mix of grades (e.g., 9, 10, and 11) continues to be a major problem. School and district coordinators have requested options such as using noninstructional days for testing, relief from instructional hour limits, and allowing testing on Saturday. The last request persists despite CDE explanations that the California Education Code does not allow schools to mandate Saturday attendance.

Test Administration Support

Test administration support included the option of precoding identification on answer documents, delivery of materials, and hotline support from AIR and NCS. Comments from all sources were overwhelmingly positive. About 75% of the respondents to our survey reported taking advantage of precoded answer documents, and the same number said they would use the option again. One school coordinator considered CAHSEE the easiest to administer of all statewide tests the school conducts (excluding logistics).

Accommodations and Regulations

Two of the observed schools did not provide any accommodations for English learners (EL) or students with disabilities. One of those two schools encouraged special education students to opt out of CAHSEE, and the other tested all students without regard to status. The only school that gave some type of accommodation to special education students grouped the students with their regular classes in their regular rooms, which allowed the proctor to give special attention to instructions. The special education students did not need extra time; in fact, their biggest problem seemed to be maintaining effort through the session. After 1 hour, most had finished and all but one had finished after 1 hour and 15 minutes. In contrast, fewer than 10% of students in a regular session were finished after 1 hour, and most took more than 90 minutes.

Although two of the observed schools had high populations of Spanish speaking students, neither school offered the option of using glossaries. In fact, there were no official glossaries for the 2001 administration since the regulations permitting glossaries had not been finalized. There was a place on the answer sheet to indicate that glossaries were provided and apparently some form of glossary was provided to a few students (as indicated by the survey). Similarly, regulations regarding calculators were not yet finalized. There was no place on the answer sheet to indicate that calculators were provided, but seven testing coordinators responding to our survey indicated calculator use.

The surveys also reflected a low frequency of accommodation. School site coordinators reported 16 cases in which special education students took advantage of calculators, glossaries, readers, or large-format materials. Because some district coordinators in the focus group raised the possibility that students in large schools might have more access to accommodation than others, the distribution of accommodations by school size is shown in Table 2.3. Although the number of accommodations is too small for any final conclusion, the percentage of schools offering some accommodation in the sample is virtually the same for small schools (45%) as for large schools (47%).

TABLE 2.3 Accommodation for Students With Disabilities by School Size *

Accommodation	Enrollment:	501+	100-500	1-99	Total
	Number of Schools:	17	14	11	42
Calculator		4	0	3	7
Glossary		0	1**	0	1
Reader		3**	2	2	7
Large Format		1	0	0	1

* Based on our Spring 2001 survey of 42 test coordinators in our longitudinal study sample. Note that policy regarding allowable accommodations was changed significantly subsequent to the 2001 administration.

** Also for EL (English learners)

Table 2.4 shows the number of students who were provided various accommodations according to information recorded on the student answer sheets. *Scheduling* accommodations generally indicated additional breaks, since all students were to be allowed almost unlimited time. This was clearly the most frequent accommodation. *Presentation*, the next most frequent accommodation, generally indicated large format text.

Accommodations for EL were even less frequent. As shown in Table 2.3 above, only one school in the survey offered glossaries to EL students and one provided the option of a reader. Coordinators were asked to identify other accommodations. These included separate rooms (two special education; one EL), extended time (three special education), and a bilingual aide (EL).

TABLE 2.4 Accommodations Reported for All Students Testing in March 2001

	ELA		Mathematics	
Accommodation	Number	Percent	Number	Percent
Scheduling	6,712	1.92	6,403	1.85
Presentation	1,530	0.44	880	0.25
Braille	108	0.03	40	0.01
Response	924	0.26	1102	0.32
Glossary	403	0.12	118	0.03
Test Read Aloud	N/A	N/A	1564	0.45

The relatively low level of accommodation was no doubt affected by uncertainty about whether results would count for graduation, which may have led to reduced participation of special education and EL students. About 40% of the surveyed coordinators reported that they tested fewer than half of the eligible students with disabilities and about 30% of EL students. In addition, coordinators in the focus group reported confusion about which means of accommodation were available. Consistent with those reports, about 40% of the school coordinators expected more accommodation in 2002.

Clearly, it would be highly desirable to ensure greater consistency in the provision of testing accommodations in future administrations. As noted below, there has been considerable discussion of accommodation policies by the SBE and CDE has conducted workshops for district test coordinators on test accommodation.

Subsequent Actions by CDE

A number of steps to further improve administration procedures have been taken since the 2001 administration. The transition to a new test developer in 2001 has included substantial coordination to improve the already high quality of workshops and test administration support. In addition, CDE has implemented policies that should ensure adequate time for administration of the ELA section and enable more comprehensive provision of accommodations. A summary of some of the more salient changes is provided here.

Adequate Time for ELA

One reason that ELA time requirements were so severe was that the ratio of items to reading passages was low, in some cases requiring students to read several paragraphs to answer just two questions. ETS recommended that additional items be developed for use in the 2002 tests, including additional items for each reading comprehension passage that had already been field-tested. ETS staff wrote the items and conducted content review and bias review panels on them. Besides reducing the time for ELA, the reviews included extensive editing of the passages, with the goal of improving their quality and enhancing the educators'

support for the ELA test. The revised passages and associated items and writing prompts will be assessed in field tests in January 2002.

The major decision in addressing ELA time requirements was to require that schools conduct the ELA part of CAHSEE over two adjacent days. Students will answer half of the multiple-choice questions and write one essay on each day. This change should greatly reduce fatigue for all students and ensure that additional time is available on the test day for students who need it. It is an aggressive, appropriate response to feedback from the field.

Although the 2-day ELA should solve fatigue problems, it could have additional unintended consequences. We are concerned that the way the decision is implemented may have an undesired impact by identifying students as "not passed" who might better be classified as "not tested" due to absence on one of the two testing days. Students who take only half of the ELA items cannot pass. Students who are absent for the first day will probably not be tested on the second day and can be readily scheduled for the next testing session about two months later. The problem is with students who take the first half of the test but are absent for the second day. If these students are considered to have “taken” the test, they may be forced to wait until test results are returned before scheduling a make-up session. This will likely cause them to miss the next testing opportunity. Besides the overriding consideration of fairness to the affected students, treating half a test as a complete test will also distort data for tracking performance for any evaluation, including potential inclusion in the Academic Performance Index. This issue is currently under review by the CDE.

Accommodations

Staff from CDE has devoted substantial resources to developing and publicizing guidance on the scope of allowed accommodations. The approved regulations identify categories of allowed accommodations, if they are specified in the student's IEP or 504 Plan. Four categories of accommodations are allowed: presentation (e.g., large print); response (e.g., transcriber); setting (e.g., individual carrel); and timing/scheduling (e.g., more frequent breaks).

The regulations also identify accommodations that are not allowed: calculators on the math portion and audio or oral presentation on the ELA portion. For some students, schools may administer the test using "not-allowed" accommodations, in which case the aid becomes a modification that invalidates the test results. However, if the student receives a score equivalent to passing the relevant part of the test with a modification, the district may petition to waive the CAHSEE requirement. Although the "waiver" process is covered in the training materials, schools are likely be confused about the policy, because allowing a test to be administrated with an invalidating modification is not a common practice.

CDE conducted workshops for special education coordinators. Because of the impact on test logistics, CDE also conducted three regional workshops for district test coordinators and special education lead coordinators. Part of the workshop included time to discuss logistical requirements. HumRRO observed the staff of a large urban district as it went through the process of identifying other teachers who needed to be included in the decisions, established a tentative date for the orientation, and developed a rough agenda for the orientation. After

the workshops, CDE distributed an extensive CAHSEE Accommodations Training Manual through district and county superintendents to each school site.

The work on finalizing and distributing regulations means that 2002 will provide the first opportunity to observe the impact of accommodations on test administration and test results. It will be important that the specific accommodations provided to a student be recorded accurately, together with the conditions justifying these accommodations, so that results can be analyzed appropriately. Further, it will be critical to identify any modifications that invalidate the test results and to flag score reports clearly if such modifications are used.

Review of Essay Scoring Procedures

HumRRO staff observed training of the table leaders and then the individual judges who rated the responses to each of the two essay questions. Briefly, the scoring process worked as follows:

- Two different judges independently scored each essay on a 0 to 4 scale. Blank or unreadable responses were flagged as unscorable.

- If the judges both agreed that the paper was unscorable or if they both gave scores and these scores did not differ by more than 1 point then the final score was the average of the two judges' ratings (or 0 if they both agreed the response was unscorable). Differences of 1 point were expected for papers near the boundary of the scoring levels ("fence sitters").

- If the judges disagreed as to whether the response was scorable, or if they gave scores that differed by 2 or more points, the paper was read and scored by a third judge (usually the table leader). If the third judge agreed with one of the first two judges, then that rating was the final score.

- In a few instances the third judge gave a different rating than either of the first two judges, usually a rating falling between the ratings of the first two judges. In this case, a fourth judge (who was generally more experienced in the scoring process) read the paper. The fourth judge's rating, which always agreed with the ratings of one of the first three judges, was taken as the final score for the essay.

Table 2.5 shows the frequency of agreement between the first two judges and the frequency of different ways in which initial disagreements were resolved based on the essays in the March 2001 test form.

TABLE 2.5 Scoring Agreement for the Essay

	First Essay Question		Second Essay Question	
Result	Frequency	Percent	Frequency	Percent
Absolute Agreement	260,381	74.4%	226,831	64.8%
Difference of 1 Point	85,586	24.5%	115,214	32.9%
Disagreement Over Scorability	669	0.2%	508	0.2%
Scorable, but Difference > 1	2,202	0.6%	4,182	1.2%

As indicated in the above table, disagreements by 2 points or more were quite rare. The first two judges reached sufficient agreement approximately 99% of the time for the first essay and roughly 98% of the time for the second essay. Where disagreements did occur, there was a reasonable process for their resolution.

Setting the Minimum Passing Score

The Raw Score Scale

Efforts to determine the minimum performance required for passing each test focused on a student's total points, or raw score, for the form of each test used in the March 2001 administration. The primary question was how many of the maximum possible raw score points a student must obtain to pass the exam.

At the first stage of scoring, a "raw score" is computed for each student. *For mathematics*, the raw score is simply the number of questions answered correctly. *For ELA*, the raw score is a weighted combination of the number of correct answers to the multiple-choice questions and the student's scores on each of the two essays. The exact equation for ELA was:

Weighted Raw Score = .7683 * MC + 3.3750 * CR

where MC is the number of multiple-choice items (out of 82) answered correctly and CR (constructed response) is the sum of the two essay scores, each of which ranges from 0 to 4 in half-point increments (except that it is not possible to get a score of 0.5). For ELA, the weighted raw scores are rounded to whole numbers. For mathematics, the raw scores range from 0 to 80. For ELA, the maximum possible raw score is:

Maximum Raw Score = .7683 * 82 + 3.3750*8 = 90.

As with most testing programs, scores were ultimately reported on a standardized scale. Raw scores are not exactly comparable across test forms due to minor differences in the difficulty and information value of the questions in each test form. Scores on this standardized scale will be comparable across different test forms. A separate translation will be developed for each different test form mapping the raw scores into scale scores. The CAHSEE standardized score scale was a linear translation of the Rasch (one-parameter) IRT scale (see for example, van der Linden & Hambleton, 1997) developed from the March administration. It ranged from 250 to 450 with the passing level mapped onto 350. The equating procedures used to map raw scores from the May form onto this same scale are described later in this chapter.

Standard Setting Panels

The test developer negotiated a subcontract with Howard Mitzel of Pacific Metrics to conduct a standards-setting workshop using the bookmark procedure explained below. The workshop was conducted May 18–20, 2001. Two HumRRO observers attended the workshop.

CDE had arranged for 90 workshop participants, 45 each for ELA and mathematics. Most participants were classroom teachers or content specialists who had been nominated by their districts. In addition, the roster included university faculty, school and district administrators, parents, and business people. About 10 had been on the HSEE Standards Panel or Technical Advisory Committee. Almost all panelists participated in all sessions relevant to their subject matter on both days. As a whole, the panels were broadly representative of the state and, because of the nomination process, knowledgeable about the California content standards and high school curriculum. Individually, the level of commitment and effort was high.

The bookmark procedure was appropriate for the purpose of identifying a minimum passing score and was implemented faithfully. The process began with a general orientation and an opportunity for each participant to take an abbreviated form of the exam. At the orientation, Mitzel stressed the need to make decisions based on test content. He described the ordered-item booklets, one each for mathematics and ELA, which listed the test questions in order of difficulty based on the March administration. For each question, participants were to discuss what made the question more difficult than the preceding questions, with particular attention to other questions from the same content standard.

Participants next moved to rooms for their content area, where they worked in groups (tables) of five or six participants, one of whom had been trained to be a table leader. Each table appeared to follow the directed procedure for discussing the knowledge and skills required by each question. A list showing the specific content standard assessed by each item was given to the math group and several tables noted that there were easy and difficult questions for each of the content standards into which the standards are organized.

After each table had discussed each of the test questions, the entire group reconvened for training on how to place a bookmark. Each participant was to place a marker to divide two item sets: items covering material each student should know and items covering material that is "maybe not needed" to get a diploma. Mitzel emphasized the differences between the bookmark placement and number-correct scores. After the training, participants worked individually to place the marker in their ordered-item booklets.

The next day, each table received a summary of individual bookmarks for the table showing the lowest, highest, and median bookmark placement. Table members discussed the rationale for their initial bookmark placements. Following this discussion, each panelist provided a revised bookmark placement. After lunch, the revised results were presented, showing the median bookmark and range for each table, along with what the pass rate would be for the median for the room. For math, many, but not all, were surprised by how low the projected pass rates were. The rate for ELA seemed to be what most participants expected. A representative from each table then described the rationale(s) for the table. Most were optimistic about the potential for students to improve during the 10^{th} and possibly 11^{th} grades. The median ratings did not change based on the impact information. One change that might be considered in future workshops would be to report the passing rates associated with the minimum and maximum bookmark placements in addition to reporting the passing rate for the median bookmark placement. That information would give participants a better understanding of the level of consensus they had achieved.

In the end, both panels recommended that the minimum passing score be set at 70% of the total possible points on each test. Though that is suspiciously close to traditional passing grades, we heard no evidence either that participants considered any criterion besides content or collaborated between content areas.

The Final Decision

CDE staff reviewed the panel's recommendations and discussed them with the superintendent. The superintendent stated that the recommendations of the standards-setting panel should be considered a long-term goal. She recommended that the provisional passing rates for initial implementation of the CAHSEE be somewhat more lenient. The specific recommendation, 60% of total possible points for ELA and 55% for math, reflected the fact that the current content standards had not been in place when members of the Class of 2004 were developing prerequisite skills. She also recommended that the State Board of Education should reexamine the test scores after students in the Class of 2004 are well into the 10th grade curriculum to determine whether students are passing in sufficient numbers to demonstrate that adequate opportunities to learn are being provided. On June 7, 2001 the SBE adopted the passing standards recommended by the superintendent.

Lack of Complete Information on the Class of 2004

The passing standard for an exam such as the CAHSEE reflects a judgment about what students *should* know and be able to do. The percentage of students who currently meet the passing standard is not a primary concern. It is customary, however, to provide standard setting panels with some information on the consequences of their recommended passing levels, specifically the expected passing rate. Anticipated passing rates are also used by the body making a final decision on the passing standards as a means of determining the reasonableness of the recommended standards.

Information on passing rates for the CAHSEE was limited for two reasons. First, students participated in the March administration on a voluntary basis and data for the students testing in May was not yet available. In addition, no information was yet available on passing rates for 10th grade students, more of whom would have completed the required curriculum. Nonetheless, the law required that 9th graders be afforded the opportunity to take and pass the exam and a substantial proportion of 9th graders (more than 70%) did choose to participate. Thus passing rates for 9th graders was a relevant statistic and, under the law, there was no opportunity to wait for 10th graders to take the exam or to obtain census testing on 9th graders.

The lack of complete census data is not a fatal flaw for the passing standards that were set. Passing information is not provided to standard setting panels until after they make initial recommendations and rarely, if ever, do they change their recommendations significantly on the basis of this information. In reaching a final decision about the recommended passing standards, CDE and the SBE had to set a policy as to who would be targeted for additional assistance and required to take the exam again. The available information on 9th grade test takers was entirely appropriate for checking the reasonableness of this policy decision.

Equating Scores from the March and May Test Forms

For a variety of reasons, it was important that students taking the CAHSEE in May be given a different test form (set of questions) than was used in the March administration. Test security was the primary reason. Even if there were no explicit compromise of test materials, test questions are frequently memorable for some students and they are likely to talk about them after the exam. Using mostly new questions on the next exam eliminates potential advantages to students who talked with those taking the first exam. In addition, the CDE wanted to release as many of the test questions as possible to illustrate the content of the exam. Using distinct test forms meant that there were more questions that could be released.

In constructing alternate forms of a test, developers always try to make each form equally difficult, as well as ensuring that each form adheres to content coverage targets and other aspects of a test blueprint. Notwithstanding their best efforts, minor differences in test difficulty are inevitably observed after each new form is administered. A whole science of test equating (see Kolen and Brennan, 1995) has evolved to control for these minor differences in test difficulty. A procedure known as an "embedded anchor" approach was used to equate scores from the May forms to the score scale based on results from the administration of the March test forms. An anchor test of 20 questions was created by reusing 20 questions from each of the March (ELA and math) tests in the May test forms.

The most important consideration in equating the May and March test forms was to estimate the expected raw score (number correct or weighted composite) on the May form for students who were right at the minimum passing level on the March form. This expected raw score was then mapped to the minimum passing point (350) on the standardized score scale. Researchers also wanted to know how students at other points on the March score scale would have performed on the May tests so that the meaning of other points, some fixed distance above or below the minimum passing level, could be maintained. We have not yet had an opportunity to review AIR's documentation of their equating analyses. Our own independent analyses are reported here.

We performed our own analyses of the test results to identify the appropriate raw-to-scale score conversion tables for the May forms. We used somewhat different statistical models, but ended up with the same results obtained by AIR to within round-off error.

As a result of the equating analyses, it was determined that a student who answered 44 of the 80 (55%) math questions correctly on the March form would be expected to answer 46 of the questions on the May form correctly. The May form of the mathematics test is slightly easier. Consequently a raw score of 46 on the May mathematics test was mapped onto a scale score of 350, the minimum passing level. The two forms of the ELA test were even more similar. A student who scored 54, the minimum for passing, on the March form would also be most likely to score 54 on the May form of the ELA test. Tables 2.6 and 2.7 show the final conversions from raw scores to the standard scale scores used for reporting for each of the 2001 ELA and mathematics test forms. These tables are based on our analyses of the final data files provided by AIR.

TABLE 2.6. Conversion from Weighted Raw Scores to Standard Scale Scores For the 2001 CAHSEE ELA Forms

Wtd. Raw Score		Scale Score	Wtd. Raw Score		Scale Score	Wtd. Raw Score		Scale Score
March	May		March	May		March	May	
F 0-7	0-7	250	29	30	310		59	361
A 8		254	30		311	60	60	363
I	8	256	31	313	312	61	61	365
L 9		259	32	32	314	62	62	367
	9	261	33	33	315	63	63	370
10		264		34	316	64	64	372
	10	266	34		317	65	65	375
11		268	35	35	318	66	66	377
	11	270		36	319	67	67	380
12		272	36		320	68	68	383
	12	273	37	37	321	69	69	385
13		276		38	322	70	70	388
	13	277	38		323	71	71	391
14		279	39	39	324		72	394
	14	280		40	325	72		395
15	15	282	40		326	73	73	398
16	16	285	41	41	327	74	74	401
			42	42	329	75	75	405
17	17	287	43	43	330		76	408
18	18	290	44	44	332	76		408
19	19	292	45	45	333		77	412
20	20	294	46	46	335	77		413
21	21	296	47	47	337		78	416
	22	297	48		338	78		417
22		298		48	339		79	420
	23	299	49	49	340	79		421
23		300	50	50	342		80	425
	24	301	51	51	344	80		426
24		302	52	52	346		81	430
	25	303	53	53	348	81		431
25	26	304	P **54**	**54**	**350***		82	435
26		305	A 55	55	352	82		437
	27	306	S 56	56	354		83	441
27	28	307	S 57	57	356	83		443
28		308	58	58	358		84	448
	29	309	59		360	84-90	85-90	450

* Scores of 350 and higher are passing scores.

TABLE 2.7. Conversion from Number Correct to Standard Scale Scores For the 2001 CAHSEE Mathematics Forms

	Raw Score		Scale		Raw Score		Scale		Raw Score		Scale
	March	May	Score		March	May	Score		March	May	Score
F	0-6	0-7	250		27	29	316			58	376
A	7		254		28	30	318		57	59	378
I		8	255		29	31	320		58	60	380
L	8	9	260		30	32	322		59		382
	9	10	264		31	33	324			61	383
		11	268			34	326		60		385
	19		269		32		327			62	386
		12	272			35	328		61	63	388
	11		273		33		329		62		390
		13	276		34	36	330			64	391
	12		277		35	37	332		63		393
		14	279		36	38	334			65	394
	13		280			39	336		64		396
		15	282		37		337			66	398
	14		283		38	40	338		65		399
		16	285		39	41	340			67	401
	15		287		40	42	342		66		402
		17	288		41	43	344			68	404
	16		289		42	44	346		67		406
		18	291		43	45	348			69	408
	17		292	P	**44**	**46**	**350***		68		409
		19	293	A	45	47	352			70	412
	18		295	S	46	48	354		69		413
		20	296	S	47	49	356		70	71	417
	19	21	298		48	50	358			72	421
	20		300		49	51	360		71		422
		22	301		50	52	362		72	73	427
	21	23	303		51	53	364		73		432
	22	24	305			54	366			74	433
	23		307		52		367		74		438
		25	308		53	55	369			75	440
	24	26	310		54	56	371		75		445
	25	27	312		55	57	373			76	448
	26	28	314		56		375		76-80	77- 80	450

* Scores of 350 and higher are passing scores.

Reporting

Results from the 2001 administration were reported at several levels. Individual score reports were provided to the students who took one or both of the tests. These reports were distributed by the schools to the students themselves and possibly also to their parents and teachers. These reports showed the student's overall scale score in comparison to the passing level of 350 and also the number and percent of questions answered correctly for each of the major content strands. For mathematics, the strands were: probability and statistics, number sense, algebra and functions, measurement and geometry, and Algebra 1. For ELA, the Reading strands were: word analysis, reading comprehension, and literary responses and analysis. The writing strands were: writing strategies and writing conventions. For ELA, the student's score on each of the two essays was shown under writing applications. A sample student report is included in Appendix A.

Aggregate reports were created for each school, district, and county, and for the state as a whole. These reports show results for all students and separately by grade, gender, race/ethnicity, language fluency, economic status, and special education program participation. For each category, the report indicates the number of students tested, the number and percent passing and failing, the average scale score, and the average percent correct for questions in each of the content strands. The ELA reports also show the average score on each of the two essays. These reports are available to the public on the CDE Web site: http://www.cde.ca.gov/ta/tg/hs/ A sample copy of a district level report is included in Appendix A.

The results by content strand in both the individual and aggregate reports provide some useful diagnostic information. Students can note areas where they have the greatest opportunity to improve and schools and districts can identify strands where their student may need more instruction. The questions for one strand may be easier or more difficult than questions for other strands, so the percent passing alone could give misleading information about a student's standing relative to other students in that area. The state-level reports do provide a basis for comparing student or school results within each strand. Appropriate comparisons to state-level results would be facilitated if the state-level results were provided on the student reports themselves.

One item that is missing from both the student and aggregate reports is any indication of measurement error. The *Standards for educational and psychological testing (AERA, APA, NCME, 1999)* include standards for score reporting. Specifically, Standard 5.10 (page 65) states:

> *Standard 5.10. When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.*

The discussion under this standard suggests "Score precision might be depicted by error bands, or likely score ranges, showing the standard error of measurement."

Interpretive information was provided on the back of the student reports that described in general terms what the tests covered and how the scores will be used. A reference to the CDE web site for more information on test content and sample questions is provided. Neither the interpretive information nor the Web site currently provide any clear information on score accuracy or measurement error.

With the possible exception of the breakout by grade, the aggregate reports provide a wide range of information about the performance of different groups of students. We note in Chapter 5 that the initial reporting by language fluency category contains some errors that are now being corrected by CDE and the development contractor. Although the reports facilitate comparisons across categories within a particular school or district, within category comparison to statewide results require users to also access the state results. Current reports could be enhanced by making it easier to compare school and district results to statewide averages.

The aggregate reports invited comparisons across schools and districts. Due to the voluntary nature of the samples of students tested in each school, results may not have been equally representative of all 9th graders in some schools. We would like to have seen a caution against inappropriate comparisons displayed more prominently in the aggregate reports.

Summary

We observed test development, administration, scoring, equating, and reporting efforts conducted by the test developer and performed our own independent analyses at several points. We did not have any significant issues with the development processes and have few suggestions for their improvement. As might be expected, given that schools and administrators received relatively short notice that these administrations of the test would count, there were several areas where test administration might be improved in future, but on the whole the process was highly successful. Similarly, the scoring and equating processes worked reasonably well and we had only minor suggestions for their improvement. Suggestions for improving the score reports include providing information about measurement error and making it easier to compare individual and aggregate results to statewide results.